Attorney Docket No. MAILP009

APPLICATION FOR UNITED STATES PATENT

STATISTICAL MESSAGE CLASSIFIER

By Inventors:

Jonathan J. Oliver
3250 Ash Street
Palo Alto, CA 94306
A citizen of Australia

Scott Roy
3250 Ash Street
Palo Alto, CA 94306
A Citizen of the United States

Scott D. Eikenberry
3250 Ash Street
Palo Alto, CA 94306
A citizen of the United States

Bryan Kim
3250 Ash Street
Palo Alto, CA 94306
A citizen of Korea

David Koblas
3250 Ash Street
Palo Alto, CA 94306
A citizen of the United States

Brian Wilson
3250 Ash Street
Palo Alto, CA 94306
A citizen of the United States

Assignee: MailFrontier, Inc.

VAN PELT AND YI, LLP
10050 N. Foothill Blvd., Suite 200
Cupertino, CA 95014
Telephone (408) 973-2585

STATISTICAL MESSAGE CLASSIFIER

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Patent Application No. _____ (Attorney Docket No. MAILP009+) entitled LEVERAGED STATISTICAL FILTERS FOR DETECTING SPAM filed July 22, 2003 which is incorporated herein by reference for all purposes.

FIELD OF THE INVENTION

The present invention relates generally to message classification. More specifically, a technique for avoiding junk messages (spam) is disclosed.

BACKGROUND OF THE INVENTION

Electronic messages have become an indispensable part of modern communication. Electronic messages such as email or instant messages are popular because they are fast, easy, and have essentially no incremental cost. Unfortunately, these advantages of electronic messages are also exploited by marketers who regularly send out unsolicited junk messages. The junk messages are referred to as "spam", and spam senders are referred to as "spammers". Spam messages are a nuisance for users. They clog people's inbox, waste system resources, often promote distasteful subjects, and sometimes sponsor outright scams.

Personalized statistical search is a technique used by some systems for detecting and blocking spam messages. Personalized statistical searches typically depend on users to sort the messages into categories. For example, the users may put spam messages into a junk folder and keep good messages in the inbox. The spam protection program periodically updates the personalized statistical searcher by processing the categorized messages. When a new message comes in, the improved statistical searcher determines whether the incoming message is spam. The updating of the personalized statistical searcher is typically done by finding the tokens and features in the messages and updating a score or probability associated with each feature or token found in the messages. There are several techniques that are applicable for computing the score or probability. For example, if "cash" occurs in 200 of 1,000 spam messages and three out of 500 non-spam messages, the spam probability associated with the word is (200/1000)/(3/500+200/1000)= 0.971. A message having a high proportion of tokens or features associated with high spam probability is likely to be a spam message.

Personalized statistical searches have been gaining popularity as a spam fighting technique because of several advantages. Once trained, the spam filter can detect a large proportion of spam effectively. Also, the filters adapt to learn the type of words and features used in both spam and non-spam. Because they consider evidence of spam as well as evidence of good email, personal statistical searches yield few false positives (legitimate non-spam email that are mistakenly identified as spam). Additionally, the filters can be personalized so that a classification is tailored for the individual. However, personalized statistical searchers also have several disadvantages. Since their training

requires messages that are categorized by the users, they are typically deployed on the client, and are not well suited for server deployment. Also, classifying email messages manually is a labor intensive process, therefore is not suitable for deployment at the corporate level where large amounts of messages are received. It would be desirable to have statistical searches that do not depend on manual classification by users, and are suitable for server deployment and corporate level deployment.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be readily understood by the following detailed description in conjunction with the accompanying drawings, wherein like reference numerals designate like structural elements, and in which:

Figure 1 is a block diagram illustrating a statistical message filter embodiment.

Figure 2 is a flowchart illustrating the processing of a message by a system embodiment that includes a statistical classifier.

Figure 3 is a flowchart illustrating the processing of a message by another system embodiment that includes a statistical classifier.

DETAILED DESCRIPTION

The invention can be implemented in numerous ways, including as a process, an apparatus, a system, a composition of matter, a computer readable medium such as a computer readable storage medium or a computer network wherein program instructions are sent over optical or electronic communication links. In this specification, these implementations, or any other form that the invention may take, are referred to as techniques. In general, the order of the steps of disclosed processes may be altered within the scope of the invention.

A detailed description of one or more embodiments of the invention is provided below along with accompanying figures that illustrate the principles of the invention. The invention is described in connection with such embodiments, but the invention is not limited to any embodiment. The scope of the invention is limited only by the claims and the invention encompasses numerous alternatives, modifications and equivalents. Numerous specific details are set forth in the following description in order to provide a thorough understanding of the invention. These details are provided for the purpose of example and invention may be practiced according to the claims without some or all of these specific details. For the purpose of clarity, technical material that is known in the technical fields related to the invention has not been described in detail so that the invention is not unnecessarily obscured.

An improved technique for improving a statistical message classifier is disclosed. In some embodiments, a classifier tests messages and attempts to make a classification.

If the message is classified by the classifier, information pertaining to the message is used to update the statistical message classifier. The classifier is preferably a reliable classifier such as a whitelist classifier, a collaborative fingerprinting classifier, an image analyzer, a probe account, a challenge-response classifier, or any other appropriate classifier. A reliable good classifier and a reliable junk classifier are sometimes used in some embodiments. In some embodiments, the same classifier may classify both good and junk messages. The classifiers may be machine classifiers or user-augmented classifiers.

As used herein, a message refers to an e-mail message, an instant message, a text message, and/or any other appropriate information transmitted electronically. For the sake of clarity, in the following examples, techniques used for e-mail messages are discussed in detail; however, the techniques are also applicable for any other types of messages.

Figure 1 is a block diagram illustrating a statistical message filter embodiment. Reliable classifiers 102 and 104 process incoming message 100 to make a classification. Although two reliable classifiers are shown, one or more classifiers may be used in other embodiments and the number of reliable classifiers may vary for different implementations. The message may be classified as spam, non-spam, or any other appropriate category. A statistical message classifier 106 is trained using the results from the reliable classifiers periodically or as messages are processed. The reliable classifier may directly update the statistical message classifier, or store the results in a knowledge base 108 that is then used to update the statistical message classifier.

The reliability of a classifier depends on how accurately it makes a classification. The reliable classifiers are so named because when they make a classification, the classification is reliable and the outcome of the classification is likely to be correct. It should be noted that the reliable classifiers sometimes do not make any classification of a message. For example, a reliable classifier may classify 20% of the messages it processes as spam, 10% as non-spam, and makes no judgment on the rest 70% of the messages. Of the messages that are determined to be either spam or non-spam, the probability of erroneous classification may be less than 1%. While the actual percentages and criteria may vary for different implementations, a classifier is considered to be reliable as long as it is able to in some cases make a more accurate classification than the statistical message classifier under training.

There are several types of reliable classifiers that may be applicable for statistical message filtering, including: an adaptive whitelist that reliably classifies non-spam messages, a collaborative fingerprinting filter that classifies spam messages, an image analyzer that is capable of determining flesh tones in pornographic spam messages, a probe account that does not belong to any legitimate user and presumably only receives spam messages, a challenge-response classifier, etc. Once a classification is made by the reliable classifier, the statistical message classifier is updated accordingly. In some embodiments, the statistical message classifier includes a knowledge base that tracks the spam probability of features in classified messages. The features may include words, tokens, message identifier, message protocol, address, hypertext markup language

document (HTML) properties or any other appropriate aspects of the message that can be used to train the statistical message classifier.

The reliable classifiers may update the statistical message classifier by processing messages such as previously stored messages, outgoing messages and incoming messages. The reliable classifiers are preferably machine classifiers that can process large amounts of messages more efficiently than manually classifying the messages. Using machine classifiers makes a statistical message classifier more suitable for server and corporate level deployment.

Figure 2 is a flowchart illustrating the processing of a message by a system embodiment that includes a statistical classifier. Once a message is received (200), it is tested with a machine classifier (202). The machine classifier is preferably a reliable one although other classifiers may also be used. The classifier attempts to classify the message and provides a classification result (204). If the message is classified as either good or spam, the statistical classifier is updated (206). If, however, the machine classifier does not make a judgment on the message, the message is then further processed (208). In some embodiments, the message is delivered to the user. In some embodiments, the message is further classified by other classifiers. In some embodiments, the statistical classifier is used to further test the message.

The techniques may be used to update a statistical message classifier for an individual user or a group of users. In some embodiments, the users share a statistical message classifier that is updated when a reliable classifier classifies the message. In

some embodiments, the users have their own statistical message classifiers. Once a reliable classifier classifies the message, the statistical message classifiers of the individual users are updated.

Figure 3 is a flowchart illustrating the processing of a message by another system embodiment that includes a statistical classifier. Once a message is received (300), it is first tested with a reliable good classifier (302). The reliable good classifier is able to make a classification of messages that are good (i.e., non-spam) reliably. In one embodiment the reliable good classifier is a whitelist classifier that classifies the message based on a database of known allowable sender addresses. The testing result may indicate that the message is good, and control is transferred from 304 to 318, where the good message is processed accordingly; in some embodiments the message is delivered to the intended recipient. If, however, the reliable good classifier makes no judgment on whether the message is good, control is transferred from 304 to 306, where the message is further tested with a reliable junk classifier. Although the reliable good classifier and the reliable junk classifier are two distinct classifiers in this embodiment, a single classifier may function both as the reliable good classifier and the reliable junk classifier.

The reliable junk classifier, for example, a classifier that uses a collaborative fingerprinting technique, is capable of reliably determining whether a message is junk. If the message is determined to be junk, control is transferred from 308 to 320 where the junk message is processed accordingly. In some embodiments, the junk message is quarantined; in some embodiments the junk message is deleted. If, however, the reliable junk classifier is unable to determine whether the message is junk, control is then

optionally transferred from 308 to 310, where other classification techniques are applied. In some embodiments, the statistical classifier is used to further test the message. If the other classification techniques determine that the message is a good message, control is then transferred from 312 to 318 where the good message is processed as such. If the message is classified as junk, control is transferred from 312 to 320, where the junk message is processed accordingly. Whether the message is determined to be good or junk, this information is useful for updating the statistical classifier. Thus, control is transferred to updating the statistical classifier (322) from both 318 and 320. The order of testing may be different for other embodiments. Although the reliable classifiers are preferably machine classifiers, the process is also applicable to classifications done by a person.

There are several ways to update the statistical classifier. In some embodiments, a training set is updated using the tokens or features of the classified message. In some embodiments, a statistical model used by the classifier is updated to reflect the classification information derived from the message. In some embodiments, in order to protect the privacy of email recipients, the information pertaining to the messages is encrypted. In some embodiments, the encryption is omitted since the tokens or features in the messages are parsed and stored in such a way that the original message cannot be easily reconstructed and thus does not pose a serious threat to privacy.

An example is shown below to illustrate how the statistical model is improved using classified messages. The reliable classifiers classify received messages and provide the statistical message classifier with a knowledge base. A message is parsed to obtain

various features. If the message is determined to be good, the "good count" for each of the features in the message is incremented, and if the message is determined to be spam, the "spam count" for each of the features in the message is decremented. Table 1 is used in some embodiments to store various features and the number of times they are determined either as good or spam:

Feature Name	Good Count	Spam Count
mortgage	10	1
auto	1	10
greeting	3	1
...	...	...

Table 1

In some embodiments, user inputs are used to augment the classification made by the reliable classifiers. Since the user's decisions are ultimately the most reliable classification available, the user-augmented classification is given extra weight in some embodiments. Table 2 is used in some embodiments to track the user classification. If a non-spam or unclassified message delivered to a user is determined to be junk by the user, the junk count is then incremented. If the message is determined to be junk by the classifier, but the user reverses the decision, the unjunk count is then incremented. Optionally, a whitelist counter is used to track the number of times a feature has appeared in whitelisted emails. Typically, a whitelisted email is email that comes from an address stored in the recipient's address book or an address to which the recipient has previously

sent a message. Instead of scoring all the whitelisted messages, in some embodiments a portion of whitelisted messages are processed.

Feature Name	Junk Count	Unjunk Count	Whitelist Count	Score
mortgage	552	7	7	-3.33
auto	132	5	186	0.58
greeting	16	1	11	0.07
...	...	...	...	

Table 2

A score for each feature may be computed based on the counter values in the tables and a predetermined score function. In one embodiment, the score is computed based on counters from both tables using the following equations:

$$CountA = S1 * SpamCount + S2 * JunkCount \quad \text{(equation 1)}$$

$$CountB = S3 * GoodCount + S4 * UnjunkCount + S5 * WhiteListCount \quad \text{(equation 2)}$$

$$FeatureScore = SCORE\_FUNCTION(CountA, CountB) \quad \text{(equation 3)}$$

where S1, S2, S3, S4 and S5 are learning parameters of the system that may be adapted to minimize error rate, and SCORE_FUNCTION is a function dependent on the statistical model that is used.

In one embodiment, the learning parameters are all equal to 1 and the following score function is used:

$$SCORE\_FUNCTION = -\log(\frac{CountA + A}{TotalSpam + B}) + \log(\frac{CountB + A}{TotalGood + B})$$

(equation 4),

where TotalSpam is the total number of spam messages identified and TotalGood is the total number of good messages identified, and A and B are prior constants that may vary in different implementations. In this embodiment, A is 10 and B is 250. For example, if "cash" occurs in 200 of 1,000 spam messages and three out of 500 non-spam messages, its feature score is computed as the following:

$$SCORE\_FUNCTION = -\log(\frac{200 + 10}{1000 + 250}) + \log(\frac{3 + 10}{500 + 250}) = -2.2713$$

A technique for improving a statistical message classifier has been disclosed. In some embodiments, the statistical message classifier is updated according to message classification by a machine classifier. In some embodiments, the statistical message classifier is updated according to the classification made by one or more other type of classifiers. These techniques reduce the amount of labor required for training a statistical message classifier, and make such classifier more suitable for server deployment.

Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. It should be noted that there are

many alternative ways of implementing both the process and apparatus of the present invention. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

WHAT IS CLAIMED IS: